Building Topic Modelling on Theses Abstracts Data

Thesis Supervisors Finder for Students

MAI VU — Information Technology in Metropolia University of Applied Sciences

December 2021

AMAZING SUPERVISORS

Aarne Klemetti — Researching Lecturer in Metropolia University of Applied Sciences

Janne Kauttonen — Staff Researcher in Haaga-Helia University of Applied Sciences

0 — Content

1 — Project Background

Project's goal, dataset, and external resources.

2 — Theoretical Background

Natural Language Processing, Topic Modelling, Text Preprocessing.

3 — Data Preprocessing

Explore the data and prepare the data for running the algorithms.

4 — Implementation & Results

Run model and present results.

5 — Further Development & Conclusion

Potential improvements, proof of concept, and summary of the thesis.

1 — Project Background

Dataset

Goal

Supercomputer

THESEUS

Data from 2009 – 2020
Collected by Janne Kauttonen

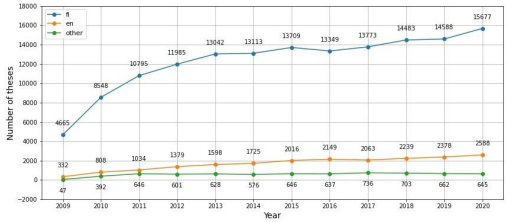
Clean data:
dropping, translating

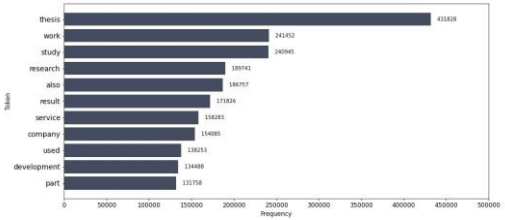
	handle	year	original_language	organization	google_translated_en	en	google_translated_fi	fi
133602	10024/149869	2018	fi	Jyväskylä University of Applied Sciences	0	The aim of the thesis was to improve the spare...	0	Opinnäytetyön tavoitteena oli tehostaa varaosa...
150857	10024/291862	2019	fi	Turku University of Applied Sciences	0	The aim of this thesis was to clarify in which...	0	Tämän opinnäytetyön tavoitteena oli selvittää,...
96648	10024/108106	2016	fi	Haaga-Helia University of Applied Sciences	1	The aim of the work was to produce material on...	0	Työn tavoitteena oli tuottaa materiaalia vuoro...

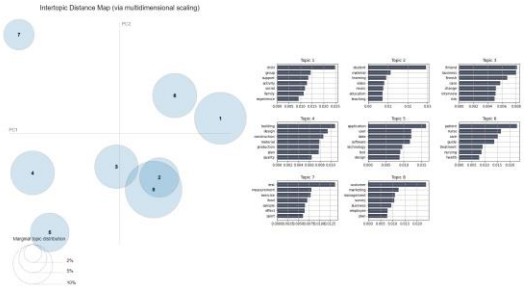
1 — Project Background

Dataset **Goal** **Supercomputer**

	handle	year	original_language	organization	google_translated_en	en	google_translated_fi	fi
133602	10024/149869	2018	fi	Jyväskylä University of Applied Sciences	0	The aim of the thesis was to improve the spare...	0	Opinnäytetyön tavoitteena oli tehostaa varaosa...
150857	10024/291862	2019	fi	Turku University of Applied Sciences	0	The aim of this thesis was to clarify in which...	0	Tämän opinnäytetyön tavoitteena oli selvittää,...
96648	10024/108106	2016	fi	Haaga-Helia University of Applied Sciences	1	The aim of the work was to produce material on...	0	Työn tavoitteena oli tuottaa materiaalia vuoro...







1 — Project Background

Dataset

Goal

Supercomputer

Deep learning approach in food detection: An application for Nutrition tracking

Loc, Hoang (2021)
Tracking food intake can give Insights into eating habits, hence it is useful to confront the rising public health Threat of obesity and overweight. Because of the high pace of modern life, people usually do not have time for ...

1

'Freedom of choice': A case study of internal communication strategies and practices on marketing in the health care sector

Pohjalainen, Sara (2021)
The planned healthcare and social services reform (SOTE) has resulted in a legislation that increases the patient's freedom of choice when deciding their health care service provider. This has opened new possibilities for ...

2

Entrepreneurs caring for the elderly: Perspectives from Finland and China : An exploratory study between Finland and China

WEI, JINQI (2021)
This thesis provides an exploratory study of the motivations and confronted challenges of entrepreneurs in the homecare sector within Finnish and Chinese contexts. Specially, through a literature review and in-depth ...

3

Machine Vision in Industrial Quality Control

Vu, Quang (2021)
With the increasing requirement of improving productivity and precision, machine vision has gained more and more traction and has continually proven to be an effective method of automated visual inspection. From food to ...

4

Source: theseus.fi

1 — Project Background

Dataset

Goal

Supercomputer

Source: Google Images

1 — Project Background

Dataset **Goal** **Supercomputer**

CSC

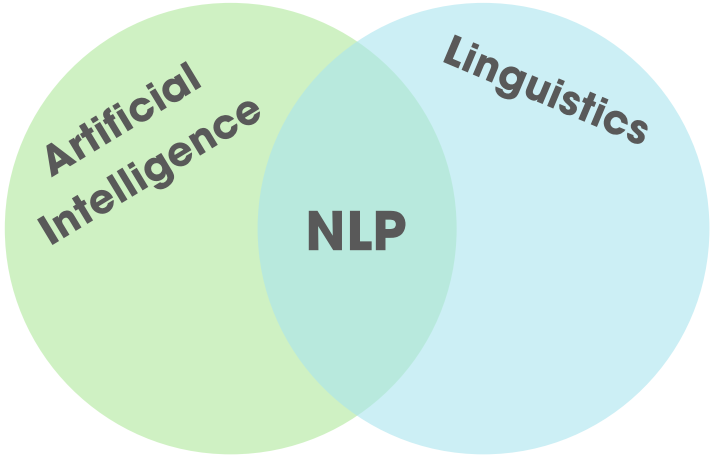
ICT Solutions for Brilliant Minds

Puhti - Machine Learning / Artificial Intelligence workloads

Longest run time: around 1 week

2 — Theoretical Background

Natural Language Processing



Topic Modeling

Q: Why can't you trust an atom?
A: Because they make up everything.

Unstructured data

Text preprocessing

Methods: Rule-based, ML, DL

2 — Theoretical Background

Natural Language Processing

Topic Modeling

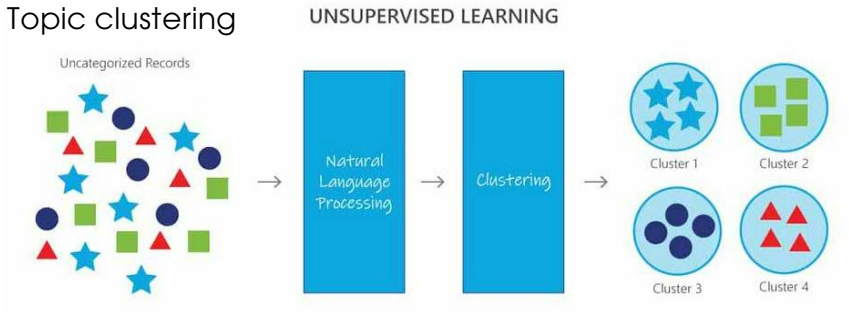
Text preprocessing

Latent Dirichlet Allocation (LDA)

Dynamic Topic Modeling (DTM)

Doc \ Topic	1	2	…
A	0.0163	0.293	…
B	0.7920	0.103	…
…	…	…	…

Topic clustering



Source: Google Images

2 — Theoretical Background

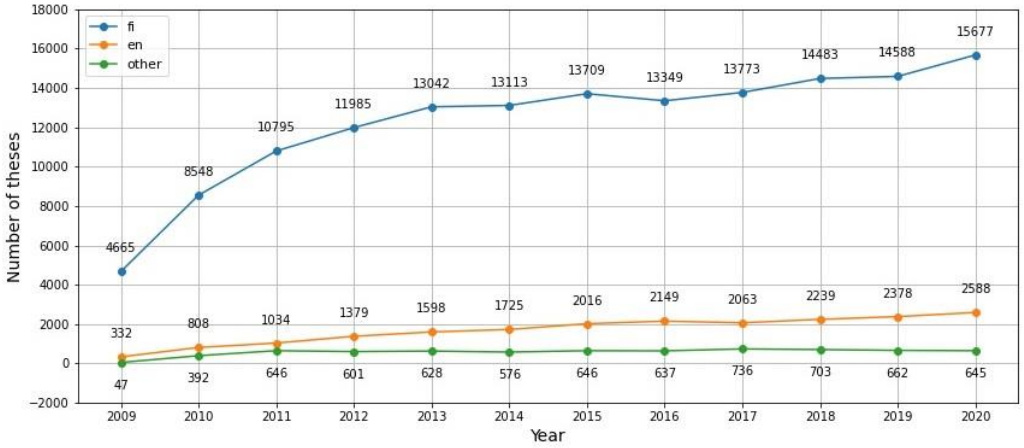
Natural Language Processing

Topic Modeling

Text preprocessing

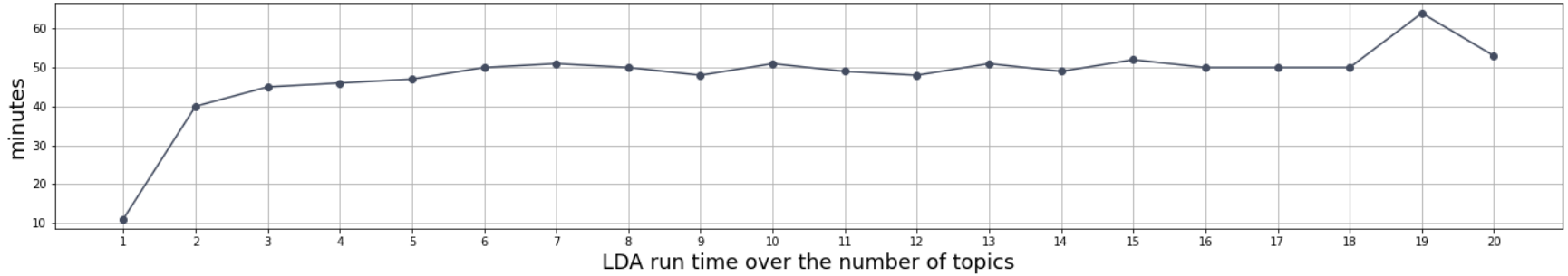
Formalization	Remove HTLM tags, punctuation, uniform texts
↓	
Tokenization	Split strings into tokens
↓	
Stop word removal	Additional dictionary, delete common words
↓	
Stemming and/or lemmatization	Simplify a word to its root
↓	
N-grams	Compound words and collocations

3 — Data Preprocessing

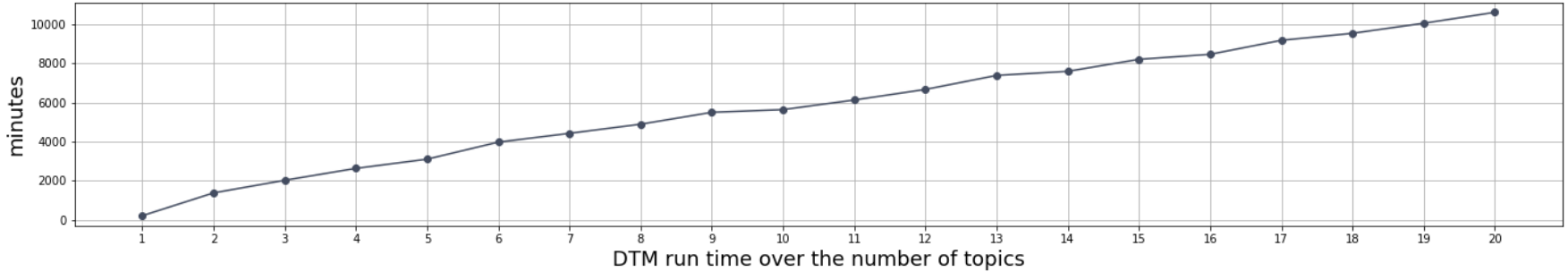


4 — Implementation & Results

LDA 8-topic model

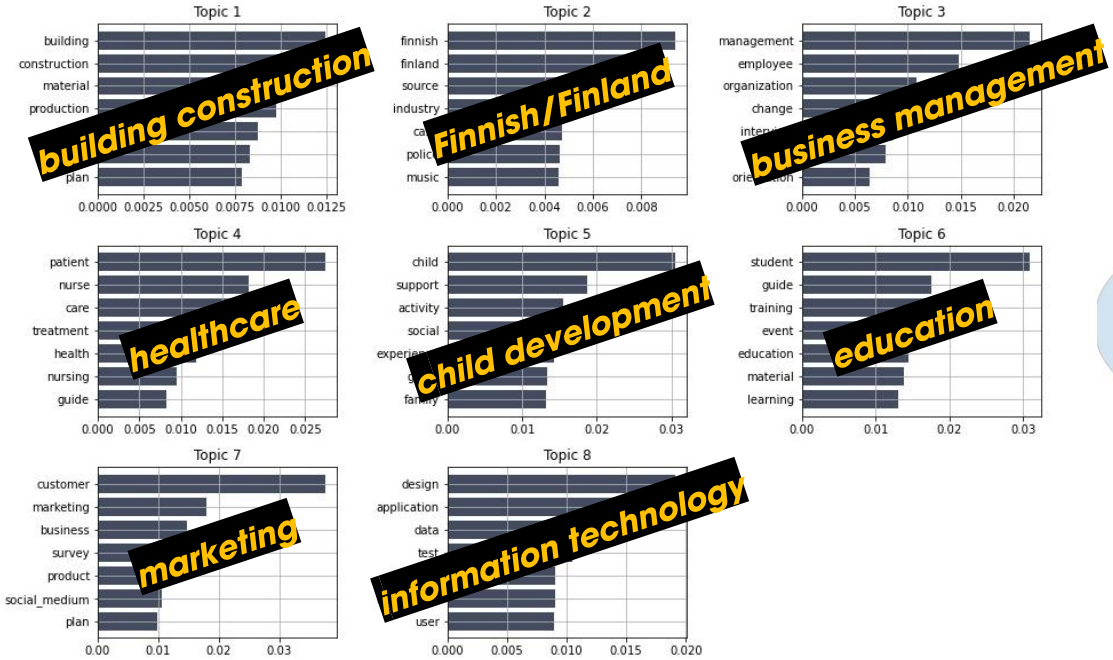


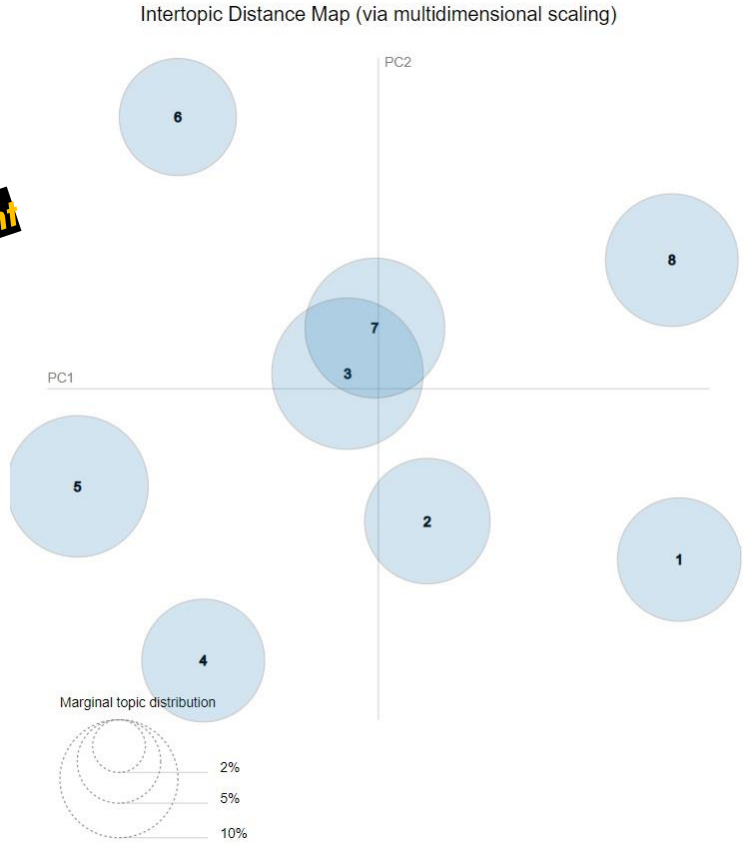
DTM 5-topic model



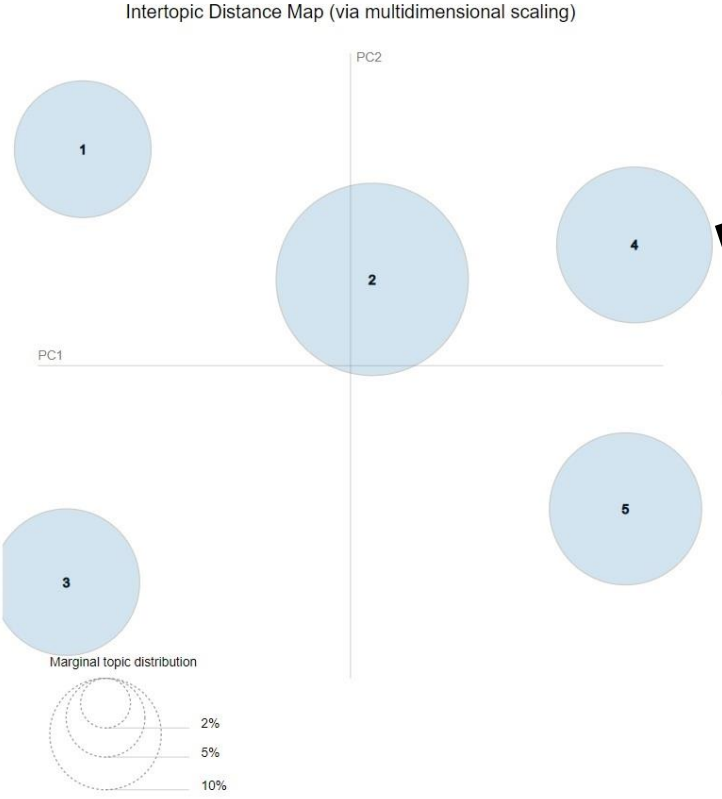
4 — Implementation & Results

LDA 8-topic model

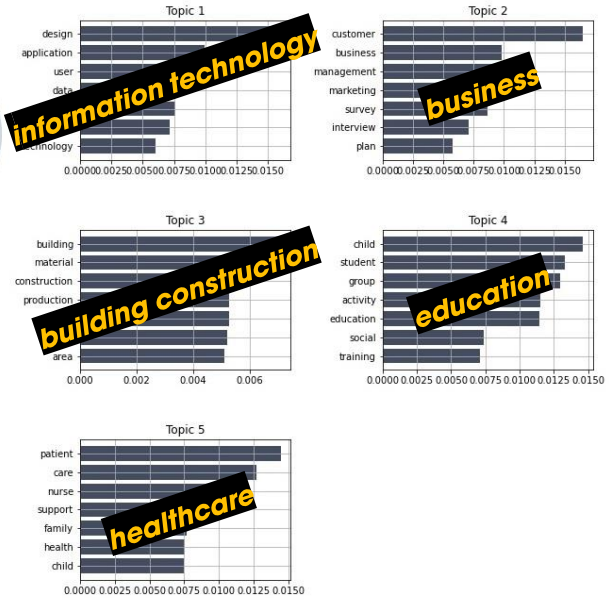




4 — Implementation & Results



DTM 5-topic model

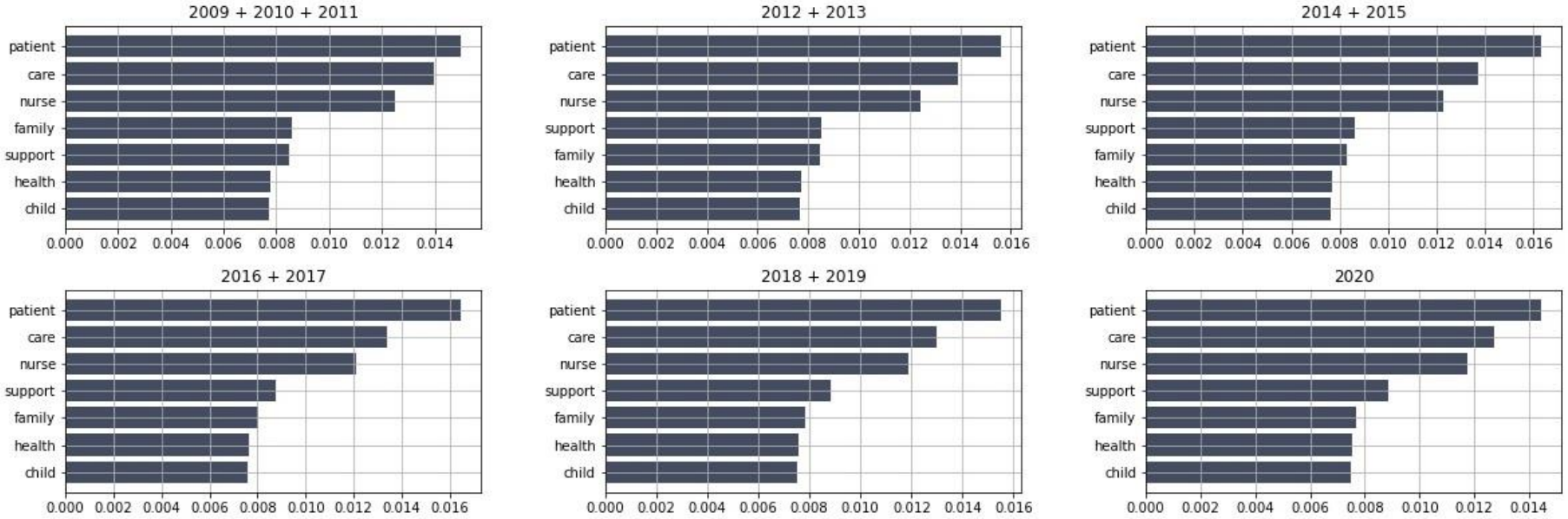


- 2009 + 2010 + 2011
- 2012 + 2013
- 2014 + 2015
- 2016 + 2017
- 2018 + 2019
- 2020

4 — Implementation & Results

LDA 8-topic model

DTM 5-topic model



Healthcare topic over time

5 — Further Development & Conclusion

Potential Improvements **Thesis Supervisors Finder** **Final Words**

Problem	Possible solution
Neutral words: *goal, people, time, change*	Better word filter
Abbreviations & bi-grams: *AI artificial_intelligence, HR human_resource, etc.*	Create a customized dictionary
Synonyms: older_people and *elderly_people*	Word embedding
Topics do not change much through time	Apply DTM for documents belong in a topic

5 — Further Development & Conclusion

Potential Improvements

"The control software is written in Python and runs on Raspberry Pi. It consists of 4 threads running simultaneously. The first and second ones are for reading weather data from the FMI Open Data service, the sun intensity data from pre-downloaded files from another FMI service called Ilmanet, and for transferring it to the simulator through REST API.

Another thread is to apply AI algorithms to calculate the setpoint and send it to the simulator. The last thread is simply for the simulator to read and update the data. Each thread runs its own loop to do the task and sleeps until its next cycle. The length of the thread's cycle can be modified as well."

Thesis Supervisors Finder

	1	2	3	4	5	6	7	8
supervisor_construction	0.604966	0.059281	0.081004	0.014471	0.014144	0.019094	0.039270	0.167771
supervisor_healthcare	0.015694	0.032550	0.066927	0.604568	0.124141	0.086014	0.019389	0.050717
supervisor_education	0.022546	0.078625	0.066176	0.079448	0.128524	0.527055	0.052795	0.044830
supervisor_it	0.116996	0.052235	0.061705	0.021204	NaN	0.034214	0.055436	0.648819
student_sample	NaN	NaN	0.033847	0.027934	NaN	0.105612	NaN	0.822165

	supervisor_construction	supervisor_healthcare	supervisor_education	supervisor_it	student_sample
supervisor_construction	0.000000	0.699431	0.657728	0.428081	0.711343
supervisor_healthcare	0.699431	0.000000	0.475373	0.682130	0.719528
supervisor_education	0.657728	0.475373	0.000000	0.644965	0.687334
supervisor_it	0.428081	0.682130	0.644965	0.000000	0.359954
student_sample	0.711343	0.719528	0.687334	0.359954	0.000000

Final Words

5 — Further Development & Conclusion

Potential Improvements

Thesis Supervisors Finder

Final Words

NLP & topic modeling

Preprocessing text data

Data analysis

Text process & word clouds

Compare the results

Potential improvements

Use a model to test

the initial idea

Building Topic Modelling on Theses Abstracts Data

Thesis Supervisors Finder for Students

MAI VU — Information Technology in Metropolia University of Applied Sciences

December 2021

AMAZING SUPERVISORS

Aarne Klemetti — Researching Lecturer in Metropolia University of Applied Sciences

Janne Kauttonen — Staff Researcher in Haaga-Helia University of Applied Sciences